BeNeLux Bioinformatics Conference – Antwerp, December 7-8 2015
Abstract ID: **000** Category: Abstract template

RANKED TILING BASED APPROACH TO DISCOVERING PATIENT SUBTYPES

Thanh Le Van[1,], Jimmy Van den Eynden[3], Dries De Maeyer[2], Ana Carolina Fierro[5], Lieven Verbeke[5], Matthijs van Leeuwen[4], Siegfried Nijssen[1,4], Luc De Raedt[1], Kathleen Marchal[5,6]*
Department of Computer Science[1], Centre of Microbial and Plant Genetics[2], KULeuven, Belgium; Department of Medical Biochemistry, University of Gothenburg[3], Sweden; Leiden Institute for Advanced Computer Science[4], Universiteit Leiden, The Netherlands; Department of Plant Biotechnology and Bioinformatics[5], Department of Information Technology, iMinds[6], Ghent University, Belgium. [*]thanh.levan@cs.kuleuven.be

Cancer is a heterogeneous disease consisting of many subtypes that usually have both shared and distinguishing mechanisms. To derive good subtypes, it is essential to have a computational model that can score their homogeneity from different angles, for example, mutated pathways and gene expression. In this paper, we introduce our ongoing work which studies a constraint-based optimisation model to discover patient subtypes as well as their perturbed pathways from mutation, transcription and interaction data. We propose a way to solve the optimisation problem based on constraint programming principles. Experiments on a TCGA breast cancer dataset demonstrate the promise of the approach.

INTRODUCTION

Discovering patient subtypes and understanding their mechanisms are essential to provide precise treatments to patients. There have been efforts to understand how mutation causes subtypes such as the work by Hofree *et al.*, (2013). However, to the best knowledge of the authors, it is still an open question on how to combine mutation and expression data to derive good subtypes. Therefore, we study a new computation model that can discover subtypes as well as their specific mutated genes and expressed genes from mutation, transcription and interaction data.

METHODS

We conjecture that a subtype consists of a number of patients who have the same set of differentially expressed genes and a set of mutated genes that hit the same pathways.

To find both mutations and expressions of patient subtypes, we extend our recent *ranked tiling* method (Le Van *et al.*, 2014). Ranked tiling is a data mining method proposed to mine regions with high average rank values in a rank matrix. In this type of matrix, each row is a complete ranking of the columns. We find that rank matrices are a good abstraction for numeric data and are useful to integrate datasets that are at different scales.

To apply the ranked tiling method, we first transform the given numeric expression matrix, where rows are expressed genes and columns are patients, into a ranked expression matrix. Then, we search for a region in the transformed matrix that has high average rank scores. However, different from the ranked tiling method, we impose a further constraint that the columns (patients) of the region should also have a number of mutated genes that have high rank scores in a network with respect to a network model. We formalise this as a constraint optimisation problem and use a constraint solver to solve it.

RESULTS & DISCUSSION

We apply our method on TCGA breast cancer dataset and discover eight subtypes. Compared to PAM50 annotations, our method divide the Basal subtype into three sub-groups named S2, S3 and S6. The LumA subtype is divided into 04 smaller groups, namely, S1, S4, S7 and S8. Finally, our method could recover the Her2 subtype in S5.

To validate the mined subtypes in the patient dimension, we assume PAM50 annotations are true labels for them. Then, grouping patients into subtypes can be seen as a multi-class prediction problem, for which we can calculate F1 score to measure the average accuracy. We also compare our scores with state-of-the-art, including iCluster+ (Mo, Q. *et al., 2013*), NBS (Hofree *et al.*, 2013) and SNF (Wang B. *et al.*, 2014). The result (not shown) illustrates that our subtypes are more homogeneous than the ones produced by iCluster+ and NBS and are comparable to those by SNF.

To validate the mined subtypes in the gene dimension, we perform hypergeometric tests to see how their mutated genes and expressed genes are related to cancer pathways. The figure below is the heatmap showing the log_10 p-values of the tests. In this Figure, we can see that the discovered subtypes have specific perturbed pathways.

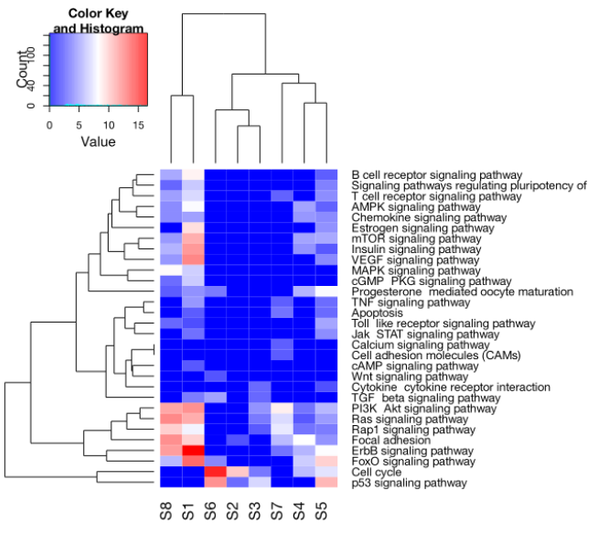

***FIGURE 1**. Cancer pathway enrichment analysis using mined mutated genes and expressed genes of subtypes*

REFERENCES

Hofree *et al.*, *Nat Methods* 10(11), 1108–15 (2013).
Le Van *et al.*, *ECML/PKDD* 2014 (2), 98–113 (2014)
Mo, Q. *et al.*, *PNAS* 110(11), 4245–50 (2013)
Wang, B. et al., *Nature methods*, 11(3), 333–7 (2014)